DOCUMENT RESUME

ED 246 109 TM 840 388

AUTHOR Younglove, William A.
TITLE A Look at Behavioristic Measurement of English Composition in United States Public Schools, 1901-1941.
PUB DATE Nov 83
NOTE 20p.; Paper presented at the Annual Meeting of the California Educational Research Association (Los Angeles, CA, November 17-18, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Historical Materials (060) -- Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Behavior Theories; *Educational History; Elementary Secondary Education; *Evaluation Methods; Grading; Measurement Techniques; Public Schools; *Rating Scales; Scaling; Scoring; Test Reliability; Test Validity; Writing (Composition); *Writing Evaluation
IDENTIFIERS Courtis (Stuart Appleton); Smith (BO); *Thorndike (Edward L)

ABSTRACT

In the early twentieth century behaviorist Edward L. Thorndike began the development and use of measurement scales to replace personal judgment to evaluate student compositions in U.S. public schools. In 1912, utilizing the Fullerton and Catell equal distance theorem, Milo B. Hillegas released the first scientifically designed scale to measure all of pupil composition. It was refined and modified over the years and numerous other scales created. Serious criticism began when scale use was at its height in 1925. Scale builder Stuart Courtis questioned scale validity, and Gestalt theorist Kurt Koffka challenged Thorndike's behavioristic psychology. In 1938 B. Othanel Smith, University of Illinois educator, thoroughly attacked the movement's logical base, and, by extension, its methods and outcomes. In that same year Courtis withdrew all his tests from the market. By 1940 there was almost no mention of composition measurement scales in the literature. Behaviorists failed to measure pupil English composition validly and reliably, which led to decline in use of Thorndikean initiated scales. However, the decades of experiment and dialogue provided increased understanding of and attention to major aspects of validity and reliability in judging composition. (ES)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED246109

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☑ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

A Look at Behavioristic Measurement of English Composition in United States Public Schools, 1901-1941

William (Bill) A. Younglove
Long Beach Unified School District

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
W. A. Younglove
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 840 388

ANNUAL MEETING OF CALIFORNIA EDUCATIONAL RESEARCH ASSOCIATION
Los Angeles Airport Hilton, Los Angeles, CA
November 17-18, 1983

ANNUAL MEETING OF CALIFORNIA EDUCATIONAL RESEARCH ASSOCIATION
Los Angeles Airport Hilton, Los Angeles, CA
November 17-18, 1983

A Look at Behavioristic Measurement of English Composition in United States Public Schools, 1901-1941

Author	William (Bill) A. Younglove
Position	teacher
Institutional Affiliation	Long Beach Unified School District
Mailing Address	3939 East Fourteenth Street Long Beach, CA 90804

The purpose of the study was to investigate early twentieth century attempts to measure pupils' behavioral responses to standard stimuli, as such responses manifested themselves in written products. Further, the study attempted analysis of the emerging problems of validity and reliability when measurement personnel, influenced primarily by Edward L. Thorndike, rendered such judgments about written composition.

Historical research of selected documents, issued between 1901-1941, provided the data from which perspective emerged. Eleven major descriptors from general and specific periodical guides supplied the base of professional periodicals and psychological and educational journals for the study. Further, some forty Thorndike authored and coauthored articles and books described the measurement methodology in detail. Additionally, seven biographical and autobiographical works on Thorndike furnished milieu.

The data revealed that behaviorists failed to measure pupil English composition validly and reliably, which led to a decline in the use of Thorndikean initiated scales. The data also showed, however, that the decades of experimentation and dialogue provided increased understanding of and attention to major aspects of validity and reliability in judging composition.

TM 840 388

A Look at Behavioristic Measurement of English Composition in United States Public Schools, 1901-1941

Personal judgment scales had been in use by teachers to rate English composition long before the turn of the twentieth century. Such scales had, for the most part, consisted of weighing the teacher count of mental errors made by pupils during writing and subtracting them from what was envisioned as a perfect paper. The score derived from such a calculus was usually interpreted in terms of a percent, which could then be translated into a letter grade. The personal scales had most likely evolved through a combination of teacher training, rhetorical tradition, and hands-on experience during the early years of teaching.

Conditions for Change

Due to a number of conditions within the schools at the turn of the century, however -- and the rise of behavioristic measurement -- the personal judgment scales were to undergo serious questioning and decline. The school conditions, not unlike those of the present, consisted of rising enrollment and attendance, particularly on the part of limited English speaking foreign-born. Sheer numbers and English composition complexity made the personal scales more difficult to use. The influence of behaviorism and the growth of measurement manifested themselves in two ways: (1) the challenge to the psychological presuppositions which underlay how the pupil produces a piece of writing and how one might determine this production, and (2) the investigation of the lack of reliability of the personal judgment method of rating compositions. This paper will touch only briefly upon the second, but will deal extensively

with the former. Suffice it to say at this point that, of fourteen major reliability studies about the ability of teachers to rate pupil English compositions, conducted between 1912-1930, only two favored teacher personal judgments over the fast rising new measurement methodology.

Behavioristic Origins

The new method of measuring English composition had its genesis in animal intelligence research conducted by Edward L. Thorndike at Teachers College, Columbia, in 1898. Thorndike became convinced that learning was the forming and strengthening of connections between situations in which the animal finds itself and impulses to action. The animal's brain neurones, excited by the senses through similarity and contiguity, formed connections, Thorndike theorized, opting toward pleasure based upon the frequency or vividness of the original stimuli.[1] Of course, Thorndike recognized the differences between the half dozen laboratory animals he had extensively worked with and human beings. Like his associationist mentor, James Cattell, however, he was drawn to the seeming simplicity of the animal's and the child's mind -- the latter, an imperfect adult, as it were. In 1901, Thorndike wrote, "Nowhere more truly than in his mental capacities is man a part of nature....Amongst the minds of animals that of man leads not as a demigod from another planet, but as a king from the same race."[2]

Early Applications

Convinced in his belief of learning through neuronal connection-formation, Thorndike turned that same year to gathering data on handwriting of children. In requesting the data on muscle

control tests of children during writing at the Horace Mann School at Teachers College,[3] Thorndike was merely following the pursuits of more than thirty other early experimenters, including such luminaries as George Fullerton, James Cattell, Charles Judd, and Robert Woodworth.[4]

In addition to gathering such data, Thorndike commenced, in 1902, to offer his Teachers College education and psychology students the first university course in educational statistics. The course announcement specified that studies would include the "means of measurement of physical, mental, and moral qualities, including the abilities involved in school subjects and rate of progress in various functions."[5]

Encouraged by the data he had gathered from his animal and human experiments and by the progress and needs of his new class, Thorndike published the first book on mental measurement, in 1904. In the preface to his Introduction to the Theory of Mental and Social Measurements, Thorndike wrote, "In the mental sciences as in the physical, we have to measure things, differences, changes and relationships or dependencies."[6]

English Composition Application

Further in the work, Thorndike speculated:

> Suppose, for instance, that we wish to find the position of a certain theme among 1,000 English themes written by first-year high-school boys. No one person can, except by accident, be a perfect rater of these, for his momentary impulse or his peculiar ideals or training will overweigh certain features. The combined opinion of ten equally good judges will always be truer than the opinion of any one of them. If, however, all the ten over-emphasized spelling or punctuation or humor, their combined rating would be false. Such a constant error in judgment is avoided as far as possible if judges are chosen at random.

> The value of having the themes arranged on a fine scale is; first, that the finer the scale, the more precise the measure, and second, that if a theme is then misplaced by chance it will not be displaced so far. For instance, if the themes were rated simply Good or Bad [sic], a theme near the dividing line, if put on the wrong side, would be put very near to the wrong side, viz., one fourth of the total distance, whereas if they were rated in 20 divisions, one in the middle would, if put to the wrong side, be moved only one fortieth of the total distance. As a practical rule one should divide the series into as many groups as one can distinguish.
>
> Amongst school abilities, achievements in handwriting, drawing, [and] ...writing English...are readily measured by serial rating, and the agreement of observers is such that great reliance can be put upon the results.[7]

What Thorndike was positing was the conception of a measurement scale based upon Fullerton and Cattell's equal difference theorem, derived from their elaborate experiments of 1892.[8] The theorem itself, based upon the normal frequency distribution curve, stated that differences that are equally often noticed are equal, unless the differences are either always noticed (would be 100%) or never noticed (would be 50%). That is to say, if 50% of a number of randomly selected, expert judges find the merit of a composition to be above a certain point and 50% find it to be equally below that same point, their judgments cancel out, and the composition may safely be said to be at that point. If, however, 75% of the judges notice that the first composition is better than a second, and the same number notice that the second is better than a third, then it can be assumed that the difference between the first and second composition is exactly the difference between the second and third, or that the difference between the first and third is exactly twice the difference between the first and second or second and third. Such a difference upon this measurement scale was designated as "1" unit of difference, or probable error.[9] The other 25% of the

judges would not see the difference between compositions whose qualities were one unit or less apart, unless trained to do so. Thorndike selected the percentages based on what he deemed were perceivable differences.

Variations in judgment were thus explained to be due to an infinite number of independent factors and that each variation was no more than an error of observation, consisting of the sum of a large number of smaller errors. The variations were believed to be qualitatively identical, each one differing from the preceding one only in amount. The equal difference theorem itself was based upon the Laplace-Gaussian, or normal frequency distribution, curve, the same curve which many accepted as being representative of the distribution of mental ability.

The Challenge to Schools

By 1905, Thorndike was calling for "trained public school men of science,"[10] a theme he was to return to often in the years ahead. In 1906, he addressed teachers themselves: "To know whether anyone has a given mental state, see if he can use it; to know whether anyone will make a given response to a certain situation, put him in the situation arranged so that that response and that response alone will produce a certain result and see if that result is produced."[11]

The First Scientific Handwriting Scale

Thorndike's continued interest in schoolchildren's handwriting and in creating scientific educational measurement merged in December 1909, when he addressed the sixty-first meeting of the American Association for the Advancement of Science on the topic of

"Units and Scales for Educational Measurement." Therein, he described the first scientifically constructed instrument for measuring the general quality of educational products. The Thorndike Handwriting Scale contained specimens of writing representing eleven degrees of excellence. Each of the equal steps in the scale was one-tenth of the difference between the best and the poorest writing of pupils in the intermediate and grammar grades. Thorndike chose general merit as the standard rather than some aspect such as legibility, beauty, or "character," because he wanted something akin to that for which school grades are oftenest given. The teacher simply needed to match the pupil's handwriting sample with its closest scale specimen. Thorndike admitted that the general merit was not some absolute, but rather was the average judgment of forty or more persons chosen at random from those deemed competent to make such judgment. He concluded, "What merit does now mean [sic] is precisely the thing measured by the scale."[12] Since the contents of the judges' minds were unknown -- and perhaps unknowable -- the validity of the judgments was assumed, rather than proven, something later critics of the scales pointed out numerous times. In other words, for Thorndike, validity was defined in terms of consistency of judgments. later to be widely accepted as a reliability factor. These two terms, their interchangeability, and their general lack of agreed upon definition compounded the communication problems among measurement experts and school lay persons alike for nearly two decades.

Nevertheless, the struggle was being made toward increased validity -- measuring what an instrument purports to measure, and increased reliability -- measuring it consistently, in the years to

come.

The First Scientific English Composition Scale

By 1912, Milo B. Hillegas, a student and associate of Thorndike, released the first scientifically designed scale to measure all of pupil composition. Culled from approximately 7,000 compositions, the final Hillegas Scale of ten divisions was derived through a series of complex steps, utilizing the Fullerton and Cattell equal difference theorem.[13] As with the Thorndike Handwriting Scale, the match between pupil samples and scale specimens was to be made apart from any consideration of age or grade of the pupil, the conditions under which the sample was written, or the purpose for which the rating was given. The ten points of the scale were, in many ways, akin to the physical scales which had been used in judging such property aspects as the hardness in minerals.

Directly after its release, the admittedly crudely built Hillegas Scale was greatly praised, particularly by measurement specialists. The scale was also introduced into various school systems, particularly through educational surveys.[14] In fact, many scientific school administrators, anxious to increase the efficiency of their systems, requested such help.

Reliability Studies, Aid Scales

A major impetus to adopt the scale came when one of Thorndike's students, Edward Elliott, teamed with Daniel Starch, to conduct a highly significant study of teacher unreliability in judging pupil composition. Their results, from a survey of 142 high schools in the North Central Association, showed teacher rater variance on the same paper to be 35-40 points on a one hundred point scale.[15]

Before long, however, recurring questions were raised about the Hillegas Scale. Initially, they had to do with the mixture of form and content within the scale.[16] Soon, questions were being raised over the ability of raters to perform the match.[17] By 1914, the aim, means, method, procedures, results, and conclusions of the Thorndike initiated scale were being questioned.[18]

Still, the scale's use continued to increase as further studies by Frederick Kelly, Baker Brownell, C.C. Certain, Nathan Silberstein, and others, during the next decade and a half, showed the shortcomings of the teacher personal judgment scales.

Scale Modification and Enlargement

When criticisms of the scale's use heightened, however, two buttressing directions were taken: (1) to modify the Hillegas Scale in accordance with the more prominent criticisms, and (2) to create new scales which attempted to eliminate the critics' delineated censures altogether.

Thorndike led the way in the first instance by extending the Hillegas Scale in 1915.[19] Other modifications included Trabue's Nassau County Supplement in 1917; Breed and Frostic's Scale, also in 1917; Willing's Scale in 1918; the Van Wagenen Minnesota English Composition Scales in 1921; and the first Hudelson Scale in 1923.

Non-Hillegas supplement scales, some of which were also determined by the consistency of expert judges, included the Rice Composition Scale, the Bliss Composition Scale, the Harvard-Newton Scales, the Courtis Gary Scale, and the Lewis Scales.

A more interesting and valuable realignment among the scales was their division into general and specific merit types.

The latter origination included the Harvard-Newton, Willing, Courtis Gary, and Van Wagenen Minnesota Scales. Their creation occurred, in part, due to the insistence of teachers that the scales provide diagnostic feedback about their pupils' writing. Researcher Baker Brownell wrote: "The ideal scale should measure a theme as definitely as a grocer's scale measures 'a pound of butter' [but][g]ood butter means color, purity, weight, consistency and so on [while] the grocer measures only weight with his scale...; [therefore,] [o]ther scales must be used for other aspects of butter."[20] The general merit scales, of course, could not consider these other aspects.

Marvin Van Wagenen of the University of Minnesota went so far as to create a calculus for obtaining a rating of general merit: $GM = \frac{4T + 2S + 1M}{7}$, wherein T was thought content, S was structure, and M was mechanics.[21] The formula was not widely adopted, however, perhaps due to a lack of identification of further writing subelements.

Leroy Sackett of the University of Texas possibly touched on an unspecified, but agreed upon, point during this period:

> The final grade is based ultimately upon the grader's subjective reaction to the composition as a whole or part by part, or in some form of relation to another sample which is different in thought content and style of presentation, or possibly to a subjective anabolic and catabolic balancing of merit and demerit attributed respectively as they are agreed upon and observed.[22]

By 1925, most of the major scales had been developed, extensively discussed, and revised. Complete works had been devoted to their use.[23] Testing conditions for teachers and pupils alike had been standardized. Fifteen major writing topics, including two for supposedly typical and maximal ability, had been formulated. At least two dozen major writing variables had been identified. The use of the

scales was at its height. As Earl Hudelson of the University of Minnesota observed, "There are already enough open-minded teachers receptive to scientific means of social control to assure the success of educational measurements, whether the rest embrace them or not...So will objective measurements come anyway...; they will divide humanity into the quick and the dead."[24]

In 1925, however, some of the quicker of the quick were beginning to point out some little noticed aspects of measurement.

Challenges to Measurement

Stuart Courtis, a scale builder, Detroit Public Schools consultant, and University of Michigan professor, indicated that the validity of the composition scales was more assumed than proven. He noted that behavior was more complex than theretofore thought to be. Responses, he pointed out, were dependent upon heredity, maturity, and a host of physical, emotional, and social factors or influences. A score on a particular test, he said, was only a measure of performance on that test; not that general ability the test may have named.[25] Courtis acknowledged his ideas contrasted with Thorndike's long-held conception of behavior.

Gestalt theorist, Kurt Koffka, also challenged behavioristic psychology that 1925 year. He was adamant that psychology could not be equated with physiological explanations and that behaviorism had excluded experimental observation and descriptive concepts. Koffka zeroed in on Thorndike's animal experiments, showing how the results observed could be interpreted differently. He argued that the animals did, at times, demonstrate learning, as C. Lloyd Morgan had experimentally proven years earlier.[26]

H. Gordon Hullfish, a year later, challenged not only the basic premises of Thorndike's behavioristic psychology, but even attempted to show a dichotomy between Thorndike's stated theoretical beliefs and his educational practice. The Ohio State researcher argued that when teachers designed the various reactions which the child should possess, they totally excluded the child. Thorndike himself, Hullfish claimed, included "ideas" as types of connection-formations, but then acted as though they were outside (determinants) of behavior.[27]

Despite these attacks, scale use continued virtually unabated through the late 1920's, when several more books were published about their use.[28] By the early 1930's, however, attacks upon the psychological base accelerated greatly. Thorndike, being greatly preoccupied with national and international aspects of objective measurement, mounted only a minor strike against the Gestalt theorists, minimizing his and their differences.[29]

Challenges by Alice Keliher of Yale, George Mead of Chicago, H. Wilbur Hatfield of the National Council of Teachers of English, David Segal of the United States Office of Education, Ralph Tyler of Chicago, Stuart Courtis of Michigan, and Barbara Burks of California came to typify what scale builder Frederick Breed called the attention gathering "radical progressive wing of the measurement movement."[30] Ralph Tyler himself opted for the term "evaluation" over "measurement" to describe the assessment instruments needed to provide for corrective feedback for writing, during the program conducted in the course of the Eight-Year Study.[31]

Attack Upon the Logical Base of Measurement

No attack upon Thorndike and the educational measurement movement, though, approached that of B. Othanel Smith. The University of Illinois educator, in 1938, attacked the movement's logical base, and, by extension, its methods and outcomes. His attack brought together the recurring elements of the past criticisms of composition measurement. He charged that: (1) experimental operations must confirm the axioms of order and amount in a scale, (2) learning is not normally distributed (He drew upon Francis Galton and David Wechsler.), (3) equal segments of the base line of a normal curve do not mark off equal units of learning (Frequencies could be laid down on the base line and there along in an order of difficulty extended, as determined by a percentage of correct solutions, but the jump from the frequencies to the units of the scale was pure inference; not verified by experiment.), (4) merit in composition is heterogeneous and complex; not well defined (Smith noted that the validity of instruments rests upon the assumption that overt responses, as manifested under the conditions of control imposed by the instrument, are fundamentally the same.), (5) variations of judgment are due not to errors of observation but to qualitative changes of judgment, (6) a response by a human is not a simple S-R occurrence, but, instead, an identification by the individual of patterns of stimuli, which the person then reconstructs, (7) only when a stimulus is fairly constant could the equal difference theorem be fairly correct, but not in the heterogeneity of merit in composition, and (8) merit is a kind of emergent quality whose character is a function of the way in which the

elements are related or organized in the observer's experience; is not a simple objective quality, but a construction.[32] In regard to the often repeated charge that scalar tests were mechanistic, Smith agreed that because they attempted to tap just one type of learning outcome, the pupil understands but does not have the attitude of understanding.

That same 1938 year, Stuart Courtis withdrew all of his tests from the market (Twenty million copies had been sold between 1909-1938.) because, he said, "I had discovered that they did not measure what they were supposed to measure."[33] By the end of that year, Courtis noted that the teachers were now more critical of the theory of measurement than the practice and that only routinized reminders of measurement occurred, principally by candidates for academic degrees.[34]

By 1940, the year of Thorndike's retirement from Columbia, almost no mention of the composition measurement scales is made in the literature. (The last are Dora V. Smith's two 1941 studies done for the state of New York and Edmiston and Gingerich's 1942 study.) In his last, entirely new book, on human nature and the social order, Thorndike repeated many of his earlier beliefs -- and attempted to extend them into new realms. For him, the human being was still an organization of S-R connections; human abilities were known from and were predictive of human behavior; and the equal difference theorem was valid.[35] In a symposium address in April of 1941, Thorndike stated, "A mental ability is a probability that certain situations will evoke certain responses, that certain tasks can be achieved, that certain mental products can be produced by the possessor of

the ability. It is defined by the situations, responses, products, and tasks, not by some inner essence."[36]

Still, the instrument which Thorndike had formulated so long ago and which had been developed to a great degree of sophistication had brought about an exploration of many variables which underlay valid and reliable judgments of pupil English composition. The decline of the use of the scales was perhaps due as much to the difficulty of the task as it was to the psychological orientation of the formulators. Certainly, it was to future researchers to recognize the errors and excesses of the past, but mainly to build upon the legacy of exploration itself to create just measurement for the written composition of pupils.

Reference Notes

1. Thorndike, E.L. Animal Intelligence. Psychological Review, June 1898, 2, 1-108.

2. Thorndike, E.L. The Evolution of the Human Intellect. Popular Science Monthly, November 1901, 60, 65.

3. Jonçich, G. The Sane Positivist: A Biography of Edward L. Thorndike. Middletown, Conn.: Wesleyan University Press, 1968, 286.

4. Thompson, M. Psychology and Pedagogy of Writing. Baltimore, Md.: Warwick and York, Inc., 1911, 49-50.

5. Teachers College, Columbia University, Announcement (1902-03), 54.

6. Thorndike, E.L. An Introduction to the Theory of Mental and Social Measurements. New York: The Science Press, 1904, 5.

7. _____. An Introduction to the Theory of Mental and Social Measurements, 19-20.

8. Fullerton, G. & Cattell, J. On the Perception of Small Differences. Philadelphia: University of Pennsylvania, 1892, 9-19.

9. Thorndike, E.L. An Introduction to the Theory of Mental and Social Measurements. New York: The Science Press, 1904, 123; 157.

10. Thorndike, E.L. The Quantitative Study of Education. Forum. March 1905, 36, 443-448.

11. Thorndike, E.L. The Principles of Teaching Based on Psychology. New York: A.G. Seiler, 1906, 260;264.

12. Thorndike, E.L. The Measurement of the Quality of Handwriting. Teachers College Record, March 1910, 11, 42.

13. Hillegas, M.B. Scale for the Measurement of Quality in English Composition by Young People. Teachers College Record, September 1912, 13, 1-60.

14. Tall, L.L. Will Standard Tests Economize Time, Lead to a Broadening of the Subject Matter, and Make for Efficiency of Both Teacher and Pupil? National Education Association Addresses and Proceedings, 1916, 64, 427.

15. Starch, D. & Elliott, E.C. Reliability of the Grading of High-School Work in English. School Review, September 1912, 20, 442-457.

16. Thomas, C.S. The Hillegas Scale. English Leaflet, January 1913, No. 104, 1-8.

17. Johnson, F.W. The Hillegas-Thorndike Scale for Measurement of Quality in English Composition by Young People. School Review, January 1913, 21, 39-49.

18. Kayfetz, I. A Critical Study of the Hillegas Composition Scale. Pedagogical Seminary, 1914, 21, 559-577.

19. Thorndike, E.L. Extension of the Hillegas Scale for the Measurement of Quality in English Composition by Young People. New York: Teachers College, Columbia University, 1915.

20. Brownell, B. A Test of the Ballou Scale of English Composition. School and Society, December 16, 1916, 4, 938-942.

21. Van Wagenen, M.J. Minnesota English Composition Scales. New York: World Book Company, 1923.

22. Sackett, L. Comparable Measures of Composition. School and Society, February 24, 1917, 5, 233-239.

23. Hudelson, E. English Composition: Its Aims, Methods and Measurements, Part I, Bloomington, Ill.: Public School Publishing Company, 1923.

24. Hudelson, E. The Effect of Objective Standards Upon Composition Teachers' Judgments. Journal of Educational Research, December 1925, 12, 329-340.

25. Courtis, S.A. Why Children Succeed. Detroit, Mich.: Courtis Standard Tests, 1925, 89.

26. Koffka, K. The Growth of the Mind: An Introduction to Child-Psychology, trans. R.M. Ogden. New York: Harcourt Brace and Company, 1925, 12-14; 153-179.

27. Hullfish, H.G. Aspects of Thorndike's Psychology in Their Relation to Educational Theory and Practice, Columbus: The Ohio State University Press, 1926, No. 1, 101-113.

28. See especially Lyman, R. Summary of Investigations Relating to Grammar, Language, and Composition. Chicago: University of Chicago, 1929.

29. Thorndike, E.L. Human Learning. New York: The Century Company, 1931, 119-131.

30. Breed, F.S. Is a Science of Education Possible? Scientific Monthly, December 1934, 39, 530-535.

31. Tyler, R.W. Defining and Measuring Objectives of Progressive Education. Educational Research Bulletin, March 18, 1936, 15, 67-72.

32. Smith, B.O. Logical Aspects of Educational Measurement. New York: Columbia University Press, 1938, vii, 55, 57, 60; 132-133; 136-150; 152-154.

33. Courtis, S.A. Forty Years of Educational Measurement: An Appraisal and a Prophecy. National Elementary School Principal. February 1946, 25, 19.

34. Courtis, S.A. Past and Present Trends in Educational Measurement. Review of Educational Research, December 1938, 8, 547-548.

35. Thorndike, E.L. Human Nature and the Social Order. New York: The Macmillan Company, 1940, 8, 105-108.

36. Thorndike, E.L. Mental Abilities. Proceedings of the American Philosophical Society, June 1941, 84, 505-508.